A Fault-tolerant Switch for Next Generation Computer Networks

V. S. Tripathi and S. Tiwari
Department of E & C E, MNNIT, Allahabad, India.
vst@mnnit.ac.in, stiwari@mnnit.ac.in

***Abstract*— In this paper, the architecture of a Multi-plane Parallel Deflection-Routed Circular Banyan (PDCB) network based switching fabric is introduced. The PDCB network has a cyclic, regular, self-routing, simple architecture and fairly good performance. Its Performance is improved due to reduction in blocking by two-dimensional path-multiplicity of the proposed architecture. It consists of 4X4 Switching Elements.**

The proposed switch is shown to be fault-tolerant. A simple analytical model based on Markov chain to evaluate the performance of proposed switch under uniform traffic condition has also been presented in this paper. The performance parameters studied are Normalized Throughput and Normalized Delay. Simulation study of the switch is also performed to validate the model proposed.

***Index Terms*— Fault-tolerant, Banyan, Markov chain, Performance, Deflection-routed, Switch.**

I. INTRODUCTION

The next generation computer networks (NGN) require high speed integrated switches[1]. Many types of switches presented in the literature are found to be suitable for the purpose[2, 3, 4, 5]. The performance parameters to compare and characterize these switches are- throughput, delay, variance of delay, buffer size, location of buffer, multicast capability, cost, fault-tolerance, reliability and scalability. Because of its regularity, modularity, self-routing property and simple architecture, a class of multistage interconnection networks called Banyan network is preferred for use in NGN[7]. It consists of 2X2 switching elements and offers same latency for all input-output pairs. It has path-uniqueness property that results in cell-sequence preservation at one hand and giving rise to the problem of blocking on the other. To solve this problem, various approaches have been proposed like- increasing the bandwidth of internal links, providing internal buffers and deflection routing[8]. Deflection routing needs multiple input-output paths within the switching fabric and adding multiplicity may further result in increase of irregularity, loss of modularity and increase in complexity [9].

The parallel plane deflection-routed circular banyan network (PDCB) based switch proposed in this paper can enhance the performance by offering a solution to the problem of blocking while maintaining the advantages of banyan architecture. This switch has self-routing property. It consists of 4X4 Switching Elements. Multiple paths are available between any input-output pair along vertical as well as horizontal axis. The packets are free to choose any free path independently. Path-multiplicity supports multicasting. It also improves the fault-tolerance and reliability of the switch. The performance of PDCB network based switch is evaluated using a new analytical model based on Markov chain and simulation. The performance parameters studied are normalized throughput, normalized delay and fault-detection coverage as they are fundamental parameters.

The remaining paper is organized as follows: Section 2 discusses the architecture of a PDCB network based switch and section 3 discusses analysis of the proposed switch. Finally the conclusion is given in section 4.

II. PDCB NETWORK BASED SWITCHING ARCHITECTURE

The deflection-routed circular banyan network is a modified banyan network that can be realized by 4X4 switching elements (SE) and augmented links as shown in figures 1 and 2. The SE in simple banyan switch is a 2X2 crossbar type network, which is connected through links to form a specific topology. The SE is modified by addition of two pairs of lateral in and lateral out ports. The additional ports are used for lateral interconnection of the SEs in a stage.

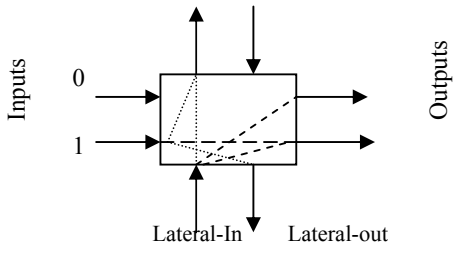

Figure 1. Modified Switching Element

In a multi-plane parallel defection-routed circular banyan (PDCB) network based switch the switching planes of figure 2 are used in parallel to make up a parallel plane switch as shown in figure 3. A large number of equivalent paths into the switching fabric exist and the packet is free to choose any free path independently. Thus switch becomes robust in the case of pathological traffic patterns. Fault-tolerance and reliability of the switch is further improved. The advantages of the PDCB switch include redundancy, augmented throughput and relaxed port speed and pin count. A switch that transmits a data packet through *k* active planes can have one or more spare planes to replace any faulty active plane.

A practical spare distribution scheme with less than 100% hardware redundancy with a three-dimensional system has been presented[11]. This methodology has been shown to be faster and more cost effective than a duplex system for a large switch since it can provide comparable availability of spares with less hardware redundancy. In our

DOI: 01.IJNS.02.01.133

switch, we adopt a similar architecture with spares in the z-axis only for redundancy purposes because this sparing scheme allows a fast (i.e., concurrent) system re-configuration.

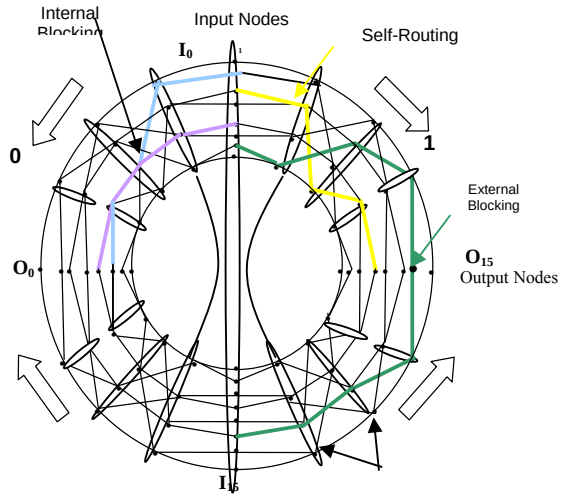

Figure 2. A 32 X 32 MCRB or DCB Network based switch

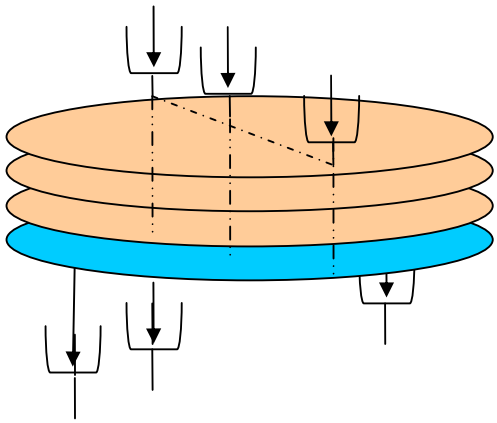

Figure 3. A Parallel Plane DCB (PDCB) switch

III. Simulation and Analysis

Exact evaluation of high speed switch performance using simple stochastic methods is very difficult[12]. The situation worsens with real-time traffic. This has led several researchers to propose approximate analysis instead of exact analysis[13]. Wu and Feng investigate a switch using a new analytical model[14]. Chen provides a queuing analysis that emphasizes delay for a switch with input port buffering using the analysis of M/G/1 queue[15]. Jenq constructs a recursive statistical model of a Banyan network with single buffers for a uniform loading. Algorithms to calculate Delay and blocking probability are also presented[16].

A Performance Evaluation

We have to make the following assumptions:

(i) Loading is balanced. The arriving packets are destined for each output with equal probability. The load on each input is $0 \le q(1) \le 1$. With a balanced load the state of each switching network in stage k should be statistically the same.

(ii) The states of the two buffers within a switching element are statistically independent. This assumption is justified by taking note that packets entering the input of a switching element originate from disjoint and independent network inputs.

We can make some definitions as follows:

$p_0(k,t)$=the probability that the switching element buffer at stage k is empty at the beginning of the t^{th} clock.

$q(k,t)$=the probability that a packet is ready to enter a switching element buffer at stage k during the t^{th} clock period.

$r(k,t)$=the probability that a packet in a switching element buffer at stage k is able to move (forward) into the next stage during the t^{th} clock period.

Now we can write a series of probabilistic equations, recursive in the stage number and in time for the above quantities:

$$q(k,t)=8/9\times p_1(k-1,t).p_1(k-1,t).p_1(k-1,t).p_1(k-1,t + 10/13 \times p_1(k-1,t).p_1(k-1,t). p_1(k-1,t).p_0(k-1,t)+ 9/12 \times p_1(k-1,t).p_1(k-1,t). p_0(k-1,t).p_0(k-1,t)+ 1/2 \times p_1(k-1,t).p_1(k-1,t).p_1(k-1,t).p_0(k-1,t); \quad k=2, 3, 4\ldots\ldots\ldots n \quad \ldots\ldots\ldots(1)$$

$$r(k,t)=[p_0(k,t)+8/9p_1(k,t)]\times[p_0(k+1,t)+p_1(k+1,t)\, r(k+1,t)], \quad k=1,2,3,\ldots\ldots,n\text{-}1 \quad r(n,t)=[p_0(n,t)+8/9 . p_1(n,t)] \quad \ldots\ldots\ldots(2)$$

$$p_0(k,t+1)=[1-q(k,t)]\,[p_0(k,t)+p_1(k,t)r(k,t)], \quad p_1(k,t+1)=1-p_0(k,t+1) \quad \ldots\ldots\ldots(3)$$

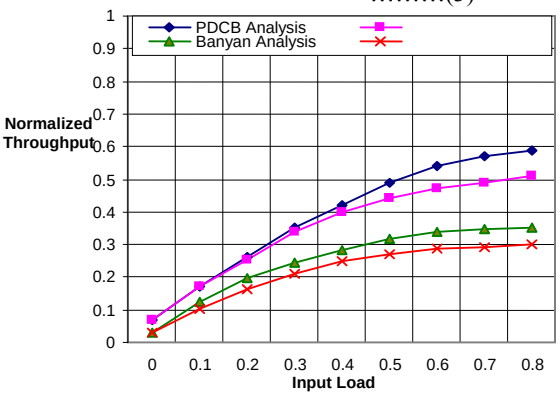

Figure 4. Plot of Normalized Throughput v/s Load

DOI: 01.IJNS.02.01.133

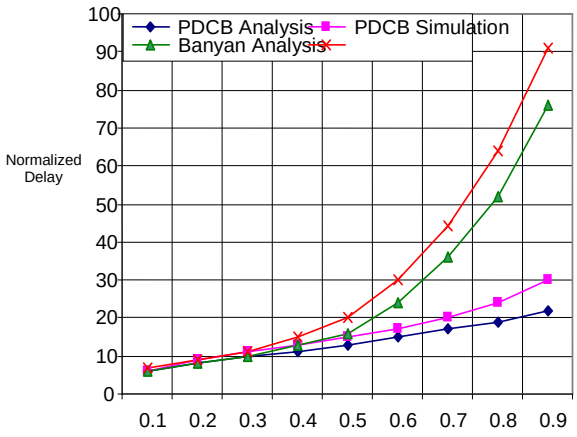

Figure 5. Plot of Normalized Delay v/s Load

The above equations can be solved iteratively for the equilibrium values. The two performance measures of most interest, as usual, are throughput and delay. The normalized throughput, $\overline{Y}$ or the average number of output packets per output link per slot is

$$\overline{Y} = p\,(k)\,r\,(k), \qquad k=1, 2, 3,\ldots\ldots, n \qquad \ldots\ldots\ldots(4)$$

The normalized average delay τ_{norm}, is

$$\tau_{norm} = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{r(k)} \qquad \ldots\ldots\ldots(5)$$

B Fault Tolerance Analysis

To study the fault tolerance property of such a fabric, we first define a fault tolerance model, fault tolerance criterion and fault tolerance method. The fault tolerance model characterizes all faults assumed to occur, stating the failure nodes (if any) for each component of the fabric. The fault tolerance criterion is the condition that must be met for the fabric to be said to have tolerated a given fault or faults.

Each link coming in or going out of an SE has a module associated with it. A module contains all of the control mechanism needed to route a request through the connecting link. A module is called an input (or output) module if its connecting link is an input (or output) link of an SE. A lateral in (or lateral out) module can be similarly named. An element is formed by an output module of an SE, an input module of an SE in the subsequent stage, and the link connecting them. The lateral out module of an SE, its connecting link, and the lateral in module of the connected SE also constitute an element. An element in the first stage contains an input module and its connecting link as shown in figure 3. Likewise, an output module (or a lateral out module and its associated switch) together with the connecting link in the last stage also form an element. An input (output) element in stage i, $0 \le i < \log_2 N-1$, is an element which contains an input (output) module of an SE in stage i. A lateral in (lateral out) element can be similarly defined. Element-based fault analysis offers better accuracy, because it is highly unlikely that a failure in an SE will disable the entire SE, especially when efforts are made in the design to avoid such an undesirable situation. So, instead of total failures, we consider partial failure of the SEs.

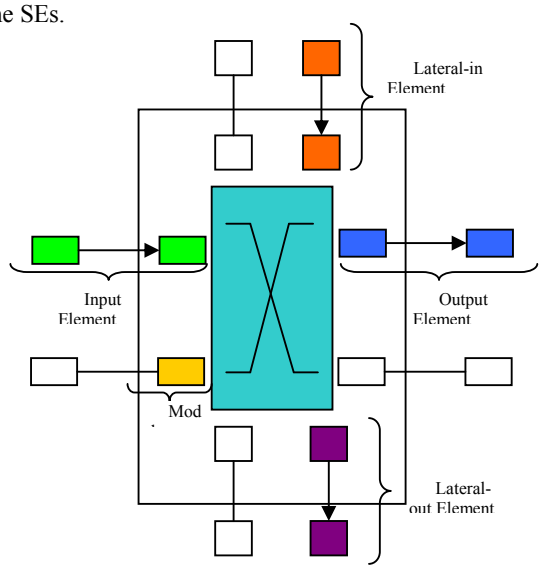

Figure 6. Switching Element used in PDCB switching fabric

1) Assumptions used

The following simplifying assumptions are made in this analysis.

- An element is considered to be faulty if any one of its components fails.
- The event that an element becomes faulty is an independent event, and it occurs randomly.
- A fabric is considered failed when the number and the locations of faulty elements prevent the connection of any path between an arbitrary input-output pair of the fabric. The faults are permanent.
- All faults are 100% detectable and can all be successfully located.
- Faults can occur at any stage except for the first stage and output links. The elements at input and output stages are highly reliable.
- The links connecting the SEs of the last stage to the output buffers are highly reliable.

An output element in stage i can be regarded as an input element in stage i +1, so we consider only output elements and lateral out elements in each stage along with the input elements of stage 0 when the total number of elements is counted. Since a PDCB switching fabric has more than one path between each input-output pair, it can tolerate at least one faulty element.

2) Fault Model

We use the fault models similar to those given by Chao[17] and Abramovici[18]. They are suitable to digital logic switching systems[18]. We have considered two faults in our model- input port controller (IPC) fault and switching

DOI: 01.IJNS.02.01.133

element (SE) fault. Some examples of the IPC faults are - Multiple Grants, Lost Grant and Grant Stuck. Some examples of SE fault models are- Idle Stuck SE, Active Stuck SE and 1/0 Stuck SE.

A PDCB fabric comprises of superimposed binary trees rooted at SE's in the first stage with SE's as their nodes and SE's in the last stage as their leaves. A path from the root to every leaf exists if all elements along the tree are fault-free. Since an input link of the first stage is shared by two external components, it is sufficient for every external component to gain access to the fabric if one half of the input elements in stage 0 are fault-free. The minimum number of fault-free elements required in a size *N* fabric can be given by:

$$N/2+(N/4-1+2)+\left(\sum_{m-1}^{n-2}2\times 2^{m}\right)+N/2= \qquad 9N/4-3$$

The total number of elements in a fabric is E = N + (3Nlog2N)/2. Hence, the maximum number of tolerable elements, denoted as BU, is {N(3.log2.N-5/2)/2+3}. An PDCB switching fabric with size 8 has BU = 29 as seen from figure 4. A conservative upper bound, denoted as B, can be obtained for PDCB switching fabric based on the following additional assumptions.

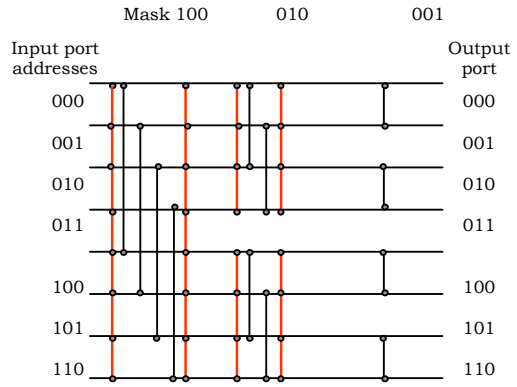

Figure 7. Fault-tolerance by multiple paths in an 8X8 PDCB switching fabric

Two faulty elements will cause the fabric to fail if

- Both faults are at the same SE and
- One of them is an output element and the other is the lateral out element.

A complete group with *M*/2 SE's has *M* output elements and *M* lateral out elements. The worst case a grouped fabric can tolerate is when all but two of the output elements in the complete group are faulty, and all but one of the lateral out elements are good. Thus, a complete group can tolerate at most (*M* - 2) + 1 = *M* - 1 faulty elements. It can be seen that stage *i*, $0 \le i \le \log_2 N-1$, has 2^i complete groups each with $N/2^{i+1}$ SE's. For stages between 0 and $\log_2 N$-2, the largest possible number of faults a fabric can tolerate is given by:

$$B= \sum_{i=0}^{\log_2 N-2} 2^i \times (N/2^i-1) = N(\log_2 N-3/2)+1.$$

3) Coverage of Fault Detection

The detection and location of the fault in the planes depend upon coverage to detect a difference of bits transmitted during the test-sequence. These bits are termed as testing bits. The testing bits are passed through all planes and a comparison of received bits reveals the fault, if any.

The fault detection coverage can be estimated by analyzing the probabilities of a single bit being 1. The probability of a legitimate bit as being of value 1is defined as $P_1(L)$ and the probability is $P_0(L)$ for a value of zero. We study the coverage when a bit collision occurs, we define $P_1(C)$ as the probability of a colliding bit (or erroneous bit) of having value "1". We also assume that $P_1(L)= P_1(C)$ and $P_0(L)= P_0(C)$

The coverage is estimated as :

$Œ = P_D/(P_D+P_U)$

Where PD is the average probability of having a detectable combination in n testing bits and PU is the average probability for an undetectable combination.

Here *i* and *j* are variables that depend on number of 1's in a combination. *k* is the number of 1's in the combination of a test sequence. ${}^nC_k$ is the number of combinations with *n* number of 1's in a combination with *k* bits.

The results obtained from above equations are shown in figure 7. Here best performance of the switch is seen for *p(1) = p(0)* = ½ and for other values of *p(t)*, performance degrades.

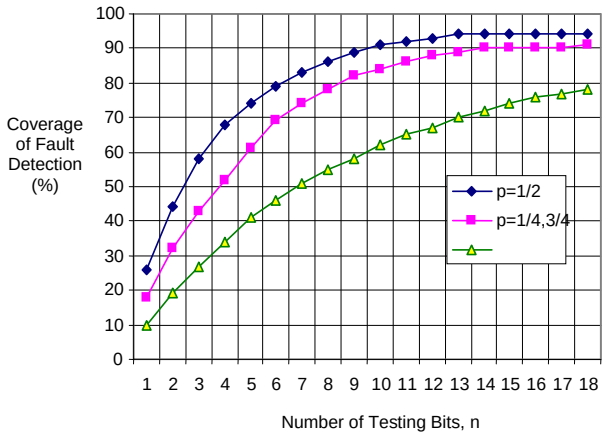

Figure 8. Coverage of fault detection

4) Terminal Reliability

Terminal Reliability (TR) is the probability that there exists at least one fault free path from a particular input port to a particular output port. TR is normally used as a measure of the robustness of a switching fabric. The set of paths in a fabric between a given input output pair is represented as a directed graph, referred to as the redundancy graph (R-graph). Generally the reliability of a system is given by $p=e^{-\lambda t}$, given a constant failure rate λ for all the components, is. However, in the fault model

DOI: 01.IJNS.02.01.133

considered in this work, the basic unit is an element, which takes the failure rate of both switching logics and the connecting link into account. Now, suppose that an element has constant reliability r. The reliability of sub-graphs observed in the switch is given by:

$\rho = r(1 - (1 - r)(1 - r2))n-2$

Now the terminal reliability between S and D can be derived by applying the bridge network result. We have, then,

$$TR_m = 2\rho r - (\rho r)^2 + 2(1-\rho)(1-r)\rho r^2(2r - r^2)$$

In order to compare we study the terminal reliability of an extra stage cube (ESC) fabric. In an ESC fabric built from 2 x 2 SE's (i.e. R = 2), a path from S (a source node) to D (a destination node) consists of n+2 elements, including the two elements for connecting S and D. The terminal reliability between S and D in an ESC fabric is given as:

$TR_{ESC} = r^2\{1- (1- rn)^2\}$

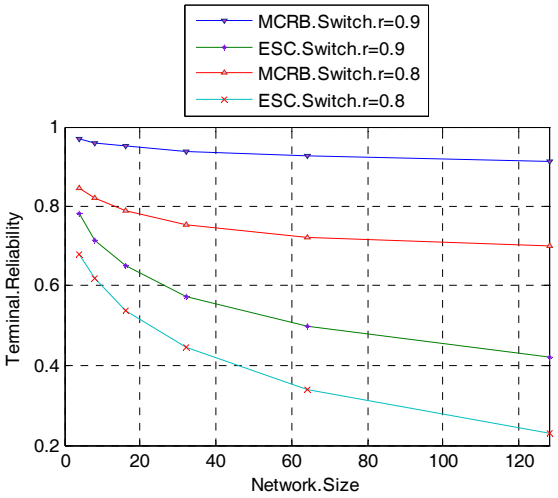

Figure 9. Terminal Reliability for different values of failure rates of two switching fabric.

CONCLUSIONS

In this paper, the PDCB switch is proposed for next generation computer networks. It is a parallel plane deflection-routed circular banyan network. The PDCB switch can deflect a packet through an alternate path along both the axes, if the desired path is not available, thereby increasing the fault-tolerance and reliability of the switch. The PDCB network has a simple self-routing scheme same as that of banyan network and it is constructed of 4 X 4 switching elements. Two pairs of input-output links in this SE are used for routing and two pairs for deflection of packets within a cyclic group of the same stage. The PDCB network has shown better throughput and delay performance because it uses the principle of controlled deflection. Thus average number of deflections is reduced and performance is improved.

REFERENCES

[1] Sadiku, N.O. Matthew, "Next Generation Network", IEEE Potentials, 2002, pp. 6-8.

[2] Yuanyang; J. Wang; "A class of multistage conference switching networks for group communication"; Proceedings International Conference on, Pages (s) : 73-80, 18-21 Aug. 2002.

[3] Tzeng, N., "Multistage-based switching fabrics for scalable routers," IEEE Transactions on Parallel and Distributed Systems, vol.15, no. 4, 2004, pp. 304-318.

[4] W. Kabacinski, G. Danilewicz; "Wide-sense non-blocking multi-log2N broadcast switching networks"; Communications, 2000. ICC 2000, 2000 IEEE International Conference on, Volume : 3, Page (s) 1440-1444 vol.3, 18-22 June 2000.

[5] E. S. H. Tse, "Switch fabric architecture analysis for a scalable bi-directionally reconfigurable IP router," J. System Architecture, vol. 50, no. 1, 2004, pp. 35-60.

[6] L. R Goke and G.J Lipovisky, "Banyan network to partitioning processor systems", Proc. 1st Annual Symp. Computer Architecture, pp 21-28, Dec. 1973.

[7] S. K. Hui, K. Seman, Kim; "An augmented chained fault-tolerant ATM switch"; Consumer Electronics, ICCE. 2002 Digest of Technical Papers. International Conference on, Page (s): 308-309, 2002.

[8] Nian-Feng Tzeng, Pen-Chung Yew, And Chuan-Qi Zhu; "Realizing Fault-Tolerant Interconnection Networks via Chaining"; IEEE Transactions On Computers, Vol. 37, No. 4. 458-462,April 1988.

[9] Cahit and J.Giglmayr, "Recirculating interconnection networks: Directed graph representations, routing and crossover minimization", 1996 Internat. Topical Meeting on Photonic in Switching, Sendai, Japan, April 1996.

[10] K. Padmanabhan, "An Efficient Architecture for Fault-Tolerant ATM Switches," IEEE/ACM Transactions on Networking, p527-537, 1995.

[11] S. S. Hussain, Y. Jenq, "Analysis and optimization of a Banyan based ATM switch by simulation", Proc. IEEE Conf. On local computer networks, pp 268, 1996.

[12] L. Kleinrock, "On the Modeling and analysis of computer networks", Proceeding of the IEEE, Vol. 81 No.8 pp. 1179-1190, August, 1993

[13] C. Wu and T.Y. Feng. "On a class of multistage interconnection networks", IEEE Trans. Computer, Aug. 1980, Page(s) 694-702, 1980.

[14] Y. C. Jenq, "On calculations of a discrete queuing system with independent general arrivals and geometric departures," IEEE Trans on Communication, Vol.28, No.6, pp. 908-910, June 1980.

[15] Y. C. Jenq, "Performance analysis of a packet switch based on a single-buffered banyan network", IEEE Journal on Selected areas in communication, Vol. SAC-1 No. 6 pp. 1014-1021, Dec, 1983.

[16] F. Lombardi, C. Feng, and W.-K. Huang, "Detection and Location of Multiple Faults in Baseline Interconnection Networks," IEEE Transactions on Computers, vol. 41, No. 10, pp. 1340-1344, October 1992.

[17] R. Rojas-Cessa, E. Oki, H. J. Chao; "Fast fault detection for a multiple-plane packet switch"; Global Telecommunications Conference,. GLOBECOM '01. IEEE Vol. 1, Page(s):102 – 109, 2001.

[18] M. Abramovici, M. A. Breuer, and A. D. Friedman, "Digital Systems Testing and Testable Design," IEEE Press, 1990.

DOI: 01.IJNS.02.01.133